УДК 004.89:004.4+004.934

А.И. Парамонов
Донецкий национальный университет, г. Донецк, Украина
anton@paramonov.info, paramonov@dongu.donetsk.ua

Интенсиональные представления в виде нечеткой гибридной модели знаний

Рассматривается проблема представления знаний, содержащихся в естественно-языковых текстах. Сформулирована проблема в общем виде и выдвинуты гипотезы на основе существующей концепции знаний. С учетом специфики человеческого мышления разработана гибридная модель представления знаний, состояние которой позволяет представлять знания, выраженные фрагментом текстовой информации. На основе модели прототипов решается задача учета индивидуальных знаний об окружении.

Информатизация жизнедеятельности человека сопровождается ростом цифровой информации и соответственно развитием систем и методов работы с ней. Особенно актуальны направления в области интеллектуальных систем и автоматизации процессов.

Одним из основных средств обмена информацией, в том числе и цифровой, между людьми является текст. Текст представляет собой связную, компактную, воспроизводимую последовательность знаков, выражающую некоторое содержание и обладающую смыслом. Принципиальной особенностью задач анализа текстовой информации является то, что предметом анализа выступают знания о предметной области, содержащиеся в текстовой информации. Иными словами, для интеллектуальных систем интерпретации текстовой информации необходимы средства для представления знаний, содержащихся в этом тексте.

Проблема понимания языковых сообщений не нова для современной науки. Решение этой проблемы лежит на стыке области искусственного интеллекта (ИИ), когнитивной психологи и лингвистики. Существуют подходы к решению данной задачи, которые описаны в работах А.Н. Хомского (теория трансформационных грамматик, теория следов), Д.А. Поспелова (представления знаний и рассуждения), Т. Винограда (вопросно-ответные системы, лингвистические процессоры), Р. Шенка (ТКЗ), М. Минского (теория фреймов), В.Н. Вагина (дедукция и вывод на сетях) и других ученых в области ИИ. В когнитивной психологии широко известны работы Дж. Андерсона, Р. Солсо, М.А. Холодной и др.

Проблема понимания естественно-языковой (ЕЯ) информации решается с помощью перевода внешнего представления на ЕЯ в некую внутреннюю структуру, описывающую систему знаний.

По мере развития исследований в области интеллектуальных систем возникла концепция знаний. Со временем сформировавшиеся модели представления знаний и данных сливались, убирая чёткую грань между данными и знаниями. В развитых моделях представления знаний можно выделить 2 компоненты [1]: интенсиональные представления и экстенсиональные представления.

Экстенсиональные представления относятся к данным. С точки зрения моделей представления текстовой информации, экстенсиональной компонентой является обрабатываемый текст.

Интенсиональные представления относятся к знаниям о предметной области. Интенсиональная компонента есть концептуальная модель представления знаний. Под концептуальной моделью понимается концепция знаний о данной области, построенная инженером по знаниям.

Автором строится гипотеза что знания, которые выражены фрагментом ЕЯ текста, могут быть представлены фрагментом концептуальной модели. Гипотезу можно сформулировать следующим образом: заданному экстенсиональному отношению, определённому как протоструктура фрагмента текста, соответствует своё интенсиональное отношение. Это главная идея, положенная в основу работы.

На основе анализа рассмотренных подходов в качестве интенсионального представления в работе предлагается гибридная нечеткая модель знаний.

Гибридная модель представления знаний

Структура гибридной модели (ГМ) разработана на основе комплексного подхода к решению поставленной задачи. Анализ когнитивного подхода позволил выделить базовые единицы модели: объекты, действия и события [2]. Элементы модели объединяются в классификационные структуры: семантические и пропозициональные сети [3], [4]. В модели учтены особенности индивидуального восприятия окружающего мира, представленные набором индивидуальных знаний о мире в виде прототипов гибрдной модели. Прототипы формируются в соответствии с классификационными структурами и разделяются на схемы и скрипты [2], [4]. Таким образом, множество элементов модели и связи между ними представляют систему знаний.

В основе модели лежит формальная система вида:

$$\text{ГМ} =< \{\text{O}\}, \{\text{D}\}, \{\text{S}\}, \text{N1}, \text{ N2} > . \qquad (1)$$

Множества {O}, {D} и {S} есть множества базовых элементов, соответственно, множества объектов, действий и событий. N1 есть классификационная структура (семантическая сеть), отражающая взаимодействия информационных единиц множеств объектов и действий. N2 есть пропозициональная сеть, отражающая связи элементов множества событий. Таким образом, будем считать предложенную ГМ интенсиональным представлением системы интерпретации текстовой информации. Каждое слово текста отражает какое-то понятие, представляемое в концептуальной схеме символом, либо его часть. Слово, или набор слов, описывающее конкретное понятие (сущность), названо термином. В предлагаемой модели понятия (термины) названы объектами.

Под объектом понимается модель сущности некоторого реального мира, которая обладает неким набором свойств. Для полноценного понимания окружения необходимо определить множество объектов рассматриваемой модели мира [5], то есть выделить предметную область. Каждый объект мировой модели может быть описан другими более элементарными объектами, а те, в свою очередь, другими и т.д.

Согласно введенному определению, объектом в концептуальной модели будем считать именованную сущность и соответствующее ей множество признаков. Формально запись объекта аналогична сетевой модели фрейма [1]:

$$O =< T_o, \{p_i\}_{i=1}^{n} >, \qquad (2)$$

где T_o – имя объекта, p_i – i-й атрибут объекта.

Отметим, что в ГМ возможно наличие различных объектов при одинаковом наборе атрибутов у них. Их отличие может быть только в имени. Возникновение такой ситуации природно для человеческого восприятия. К этому может привести недостаточное знание предметной области либо абстрагирование объектов с выделением только общих признаков.

Текст (или языковое общение), кроме рассматриваемых объектов реального мира, несет еще информацию и об отношениях между ними. В языковой грамматике такие отношения названы глаголами. Отношения между понятиями на концептуальном уровне названы действиями. За основу в определении действий была взята теория Р. Шенка, предложенная в работе [2].

По аналогии с объектами действия могут быть описаны через другие действия – более элементарные, а те, в свою очередь, через другие. Согласно теории концептуальной зависимости (ТКЗ) используется класс элементарных действий, названных АКТами [2].

Таким образом, действие какой-либо концептуализации есть, в сущности, имя, под которым объединяется множество последовательных действий, рассматриваемых как его часть. Эти инструментальные концептуализации не связаны причинно, поскольку они реально неотделимы друг от друга. Они в действительности выражают одно действие и, таким образом, рассматриваются как часть концептуализации.

Формально в работе действие концептуальной модели представлено именем и набором соответствующих ему АКТов:

$$D =< T_D, \{p_i\}_{i=1}^{m} > , \qquad (3)$$

где T_D – наименование действия, p_i – АКТ, составляющий часть действия (признак действия).

Событием названо некоторое действие, описывающее конкретные отношения конкретных объектов. В терминах предложенной модели событие есть множество, состоящее из действия и объектов, а также ряда прочих (вспомогательных) аргументов, которые могут выступать объектами или субъектами действия в элементарных АКТах. К вспомогательным аргументам относятся данные, при необходимости, дополнительно описывающие ситуацию: место, время, «инструменты» и так далее.

Итак, делается заключение, что одна концептуализация действия рассматривается как представитель одного события.

В общем виде событие описывается как упорядоченный набор S:

$$S =< O_i, D, O_j, \{c_l\}_{l=1}^{k} > , \qquad (4)$$

где O_i, O_j – объект и субъект действия (i $\neq$ j) (2), D – действие (3), c_l – дополнительный аргумент события.

Порядок объекта и субъекта действия устанавливается однозначно.

В силу того, что существуют возвратные глаголы, в которых реципиентом выступает сам субъект действия, вводится вспомогательный объект «is». Использование вспомогательного объекта возможно только для заданного множества действий.

Описанные элементы гибридной модели, объекты и действия, образуют единое пространство – знание о предметной области. Для удобного и корректного использования элементы этого пространства нуждаются в структурировании. На основе обзора существующих моделей представления семантических структур, а также с учётом представлений объектов и действий в ГМ, в качестве базовой классификационной структуры элементов ГМ выбрана модель сравнительных семантических признаков.

Элементы ГМ, объединённые в семантические сети (объекты и действия), определены как рабочие элементы гибридной модели (РЭГМ).

Семантические сети позволяют выделить все понятия в рамках предметной области и отношения между этими понятиями, а также структурировать их (разбить на составляющие и установить между всеми составляющими именованные связи).

В работе сети представлены подграфами вида:

$$N(V, A), \quad (5)$$

где V – множество узлов сетей (РЭГМ), A – множество дуг сетей.

Согласно выбранной модели сравнительных семантических признаков элементы сети (5) запишутся как (6) и (7).

$$v_i = < T, \{[P_{def}]_{k=0}^{q}\}, \{[P_{sp}]_{l=0}^{m}\} >, \forall v_i \in V, i = \overline{1, n}, \quad (6)$$

где v_i – i-й узел сети, T – имя узла семантической сети, P_{def} – определяющий признак узла, P_{sp} – характерный признак узла, q – число определяющих признаков узла, m – число характерных признаков узла.

$$A = \bigcup \{A_{type}\}_j, \quad (7)$$

где A_{type} – заданный тип связей в семантической сети.

В терминах рассматриваемой модели в качестве узлов семантических сетей будет выступать множество объектов (2) или действий (3):

$$V = \{O_i\}_{i=1}^{n} \text{ или } V = \{D_i\}_{i=1}^{n}.$$

В предложенной ГМ на семантических сетях введено два типа связей:

$$A = A_{type1} \bigcup A_{type2},$$

где A_{type1} – связи в сети типа «part_of», A_{type2} – связи в сети типа «is_a».

Для представления знаний, содержащихся в каждом отдельном предложении, используется пропозициональная репрезентация. Пропозициональный анализ представляет в памяти сложные предложения в виде простых, которые выступают абстрактными пропозициональными единицами, или эпизодами. Пропозициональная информация может быть представлена в сетях, которые показывают отношения между понятиями.

Эпизоды – это события ГМ, которые объединены пропозициональными связями. Различаются три пропозициональных связи: отношение «затем», ведущее к новому событию, отношение «во время» (или в то время как), позволяющее соединять текущие события с новым событием, и отношение «время», позволяющее соединять события с неопределенным временным порядком. Эпизоды являются узлами пропозициональной сети. Таким образом, пропозициональная сеть есть множество событий с заданными пропозициональными связями между его элементами. Пропозициональная сеть описана в гибридной модели подграфом вида:

$$N(U, B),$$

где U – множество узлов сети, B – множество дуг (пропозициональные связи).

Множество пропозициональных связей может быть записано как:

$$B = B_{after} \bigcup B_{now} \bigcup B_{time},$$

где B_{after} – множество дуг пропозициональной сети, помеченных связью «затем», B_{now} – множество связей «во время», B_{time} – множество связей «время».

Интенсиональные отношения в ГМ представлены прототипами.

Прототип – это обобщенное представление, в котором воспроизведен набор общих и детализированных признаков типичного объекта (или действия) и которое выступает в качестве основы для идентификации любого нового понятия [6]. Прототипом в ГМ принято называть любой элемент модели (объект, действие, событие), который обладает набором собственных атрибутов, доступных к изменению под заданное окружение. В работе определено два типа прототипов: схемы и скрипты.

Схемой названа абстракция объектов или репрезентация семантической сети объектов в терминах конкретного окружающего мира. Наличие и типы связей между объектами в сети зависят от модели окружающего мира. Схема отражает текущее представление заданного окружающего мира. Схемой назван подграф N', представленный в виде $\mathrm{N}'(\mathrm{V}', \mathrm{A}')$, где V' – множество узлов сетей, элементы которого представлены в виде (2) и (3), A' – множество дуг сетей ($\mathrm{A}' \subset \mathrm{R}(\mathrm{A})$).

Отличие одной модели мира от другой представлено количеством признаков в соответствующих узлах сети (значения n и m в формулах (2) и (3)), а также их составом.

Скрипты – репрезентация схемы, предложенная Шенком и Абельсоном для понятий, описывающих события [2]. В терминах предлагаемой ГМ скриптом названа репрезентация пропозициональной сети в терминах конкретного окружающего мира. Скриптом назван подграф N'', представленный в виде $\mathrm{N}''(U, B')$, где U – множество узлов сети или пропозиций, B' – подмножество возможных связей из универсума дуг пропозициональной сети ($B' \subset R(B)$).

Таким образом, скрипт задает только наличие и типы связей в сети.

С учетом введенного понятия прототипа рассматриваемая гибридная модель в конкретной реализации представляет собой семантические и пропозициональные сети прототипов. Это свойство модели демонстрирует особенность восприятия мира человеком в зависимости от его познаний.

Нечеткая гибридная модель

Предложенная и описанная выше ГМ представления знаний в виде множества прототипов (схем и скриптов) является интенсиональным представлением уникальных знаний о некоторой предметной области. Но данное представление не может быть полным, потому что человеческое мышление представляет собой нечёткий механизм [7]. Таким образом, полная модель знаний должна быть представлена нечеткой ГМ (НГМ), которая задана системой вида

$$НГМ = <\text{ГМ}, \{\text{CF}\}>,$$

где каждой составляющей ГМ из (1) приписывается фактор уверенности (CF).

Фактор уверенности для множеств объектов и действий, обозначенных в работе как РГЭМ, заложен в природе этих элементов. Исходя из (2) и (3) РЭГМ есть набор признаков, формируемый из множества всех возможных признаков при построении интенсиональной части модели. Для каждого интенсионального представления зада-

ется разная степень принадлежности определённого признака конкретному понятию. Это даёт возможность определить каждый РЭГМ как подмножество признаков, обладающее нечеткой характеристикой:

$$\underset{\sim}{T} = \{(p_i \mid \mu_{\underset{\sim}{T}}(p_i))\};\qquad(8)$$

$$P = \{p_i\};$$

$$M = \{0, \varepsilon, 2\varepsilon, .., 1\},$$

где $\mu_{\underset{\sim}{T}}(p_i)$ – функция принадлежности признака узлу $\underset{\sim}{T}$, p_i – признак РЭГМ, P – множество всех признаков, M – множество возможных значений функции принадлежностей.

Фактор уверенности для элементов множества событий задаётся уверенностью в составляющих этого события. Согласно (4) событие состоит из элементов множеств объектов и действий, а также дополнительных аргументов, которые влияют на отношения между событиями, но не являются составляющими уверенности в самом событии.

Таким образом, если в качестве элементов события принять РЭГМ, описанные в (8), то событие определяется как множество нечётких подмножеств.

$$S_j = \{\underset{\sim}{T}_i^{о}, \underset{\sim}{T}_i^{с}, \underset{\sim}{T}_i^{д}\},\qquad(9)$$

где $\underset{\sim}{T}_i^{о}, \underset{\sim}{T}_i^{с}$ – узлы семантической сети объектов (объект и субъект), $\underset{\sim}{T}_i^{д}$ – узел семантической сети действий.

Фактор уверенности классификационных структур НГМ есть множество нечётких характеристик связей в этих структурах. Соответственно, семантические сети, описанные в (5), будут представлены в виде нечёткого графа $\underset{\sim}{A}$ [7]

$$\underset{\sim}{A} = \{a_{ij} \mid \varphi(a_{ij})\},$$

$$\forall(v_i, v_j) \in V \times V : \mu_{\underset{\sim}{A}}(v_i, v_j) \in M,$$

$$\varphi(a_{ij}) = \mu_{\underset{\sim}{A}}(v_i, v_j),$$

$$M = \{0, \varepsilon, 2\varepsilon, ..., 1\},$$

где *vi*, *vj* – узлы семантической сети, *aij* – дуга из узла *vi* в узел *vj*, $\mu(v_i, v_j)$ или $\varphi(a_{ij})$ – функция принадлежности дуги *aij* данной сети, или сила связи узлов *vi* и *vj* ($\varphi(a_{ij}) \in [0,1]$), M – множество принадлежностей элементов множества $V \times V$.

С учётом формулы (8) множество узлов семантической сети $\underset{\sim}{A}$ будет представлено как множество нечётких подмножеств

$$V = \{\underset{\sim}{T}_j\}\cdot$$

Аналогичным образом нечёткое подмножество $\underset{\sim}{B}$, такое, что

$$\forall(u_i, u_j) \in U \times U : \mu_{\underset{\sim}{B}}(u_i, u_j) \in M,$$

$$M = \{0, \varepsilon, 2\varepsilon, ..., 1\},$$

задаёт нечёткий граф пропозициональной сети.

Соответственно, $\varphi(b_{ij})$ определяет силу связи узлов в данной пропозициональной сети. А сами узлы с учётом (9) будут определены как

$$U = \{S_j\},$$

где S_j – событие, или эпизод пропозициональной сети.

Построенная с учётом заданных факторов уверенности НГМ описывает индивидуальные знания о предметной области и формирует новые интенсиональные отношения.

Выводы

Исследуемая в работе проблема представления знаний сформулирована в виде гипотез на основе существующей концепции знаний. С учетом возможных интенсиональных представлений разработана гибридная модель знаний, состояние которой позволяет представлять знания, выраженные фрагментом текстовой информации. В работе показано, что необходимо учитывать индивидуальные знания об окружении. В качестве решения этой задачи предлагается модель прототипов. Но данное представление не может быть полным, потому что человеческое мышление представляет собой нечёткий механизм. Таким образом, полная модель знаний, или интенсиональное представление, в работе описана в виде нечеткой гибридной модели знаний.

Литература

1. Искусственный интеллект: В 3 кн. / Под. Ред. Д.А. Поспелова. – М.: Радио и связь, 1990. – Кн. 2 Модели и методы: Справочник. – 304 с.: ил.
2. Шенк Р. Обработка концептуальной информации: Пер. с англ. Г.В. Сенина. – М.: Энергия, 1980. – 360 с.: ил.
3. Солсо Р. Когнитивная психология. – СПб.: Питер, 2002. – 592 с.: ил.
4. Андерсон Дж. Когнитивная психология. – 5-е изд. – СПб.: Питер, 2002. – 496 с.: ил.
5. Виноград Терри. Программа, понимающая естественный язык: Пер. с англ. С.С. Калитина. – М.: Мир, 1976. – 295 с.
6. Хофман И. Активная память. – М.: Прогресс, 1986. – 39 с.
7. Кофман А. Введение в теорию нечетких множеств: Пер. с франц. – М.: Радио и связь, 1982. – 432 с.: ил.

А.І. Парамонов

Інтенсіональні представлення у вигляді нечіткої гібридної моделі знань

Розглядається проблема представлення знань, що містяться у природно-мовних текстах. Сформульована проблема в загальному вигляді і висунуті гіпотези на основі існуючої концепції знань. З урахуванням специфіки людського мислення розроблена гібридна модель представлення знань, стан якої дозволяє представляти знання, виражені фрагментом текстової інформації. На основі моделі прототипів вирішується завдання врахування індивідуальних знань про оточення.

A.L. Paramonov

Fuzzy Hybrid Model of Knowledge as Intensional Representation

The problem of knowledge representation containing in naturally-language texts is considered. The problem in a general view is formulated. Hypotheses, on the basis of the existing knowledge concept are put forward. Separation of existing models of knowledge representation is marked at the problem decision. These models do not solve all problems connected with knowledge representation. The hybrid model of knowledge representation is developed. The condition of hybrid model allows representing the knowledge expressed by a text information fragment. On the basis of prototypes model the problem is solved with consideration of individual knowledge of an environment. In view of specificity of human thinking, the indistinct hybrid model of knowledge representation is offered.

Статья поступила в редакцию 16.07.2008.